خوارزمية عامة لتوليد بيانات تخضع لتوزيع بيتا متعدد المتغيرات

د. صلاح حمزة عبد*

المستخلص

سبق في عام 2002 وقدم كل من Michael و Schucany [2] ، جهداً رائعاً ، تمثل في أسلوبهما الذي أطلقا عليه تسمية الأسلوب الخليط لتوليد بيانات تخضع لتوزيعات ثنائية المتغيرات Bivariate . سنقوم في هذا البحث بطرح خوارزمية عامة لتوليد بيانات تخضع لتوزيع بيتا متعدد المتغيرات باستخدام التعميم الذي طرحه كل من Minhajuddin و Harris و Schucany عام 2004 [3] للأسلوب الخليط المنوه عنه أعلاه ، إذ جعلوه ممكن التطبيق على التوزيعات متعددة المتغيرات .

1- المقدمة

لا يخفى على المتتبع للنظرية الإحصائية أهمية توليد المشاهدات على وفق أنموذج أو مسألة ما للأغراض التجريبية . وعلى الرغم من أن هذا المسلك قد يكون صعباً في أحيان كثيرة ، إلا انه يزداد صعوبة في حالة المتغيرات المتعددة ، بسبب الارتباط الحاصل مابين المتغيرات التي تخضع بمجملها للتوزيع المتعدد المتغيرات ، وكيفية استكشافه ، واحتسابه والتعامل معه .

إن الدراسات القائمة على المحاكاة تستند على توليد المشاهدات على وفق هذا التوزيع الاحتمالي أو ذاك لتحقيق جوانب رياضية نظرية قد تعجز الطرق التحليلية الرياضية عن تحقيقها لبعض المسائل المتعلقة بأحد شقي النظرية الإحصائية ، التقدير أو الاختبار .

إن استناد السواد الأعظم من النظرية الاحصائية على فرض التوزيع الطبيعي ، سواءً أكان أحادي أو متعدد للمتغيرات ، لا يعني أن ليس هناك ظواهر في الواقع العملي لا تخضع لهذا التوزيع ، بل قد أثبتت الدراسات بأن معظم البيانات المأخوذة من ظواهر حقيقية ، هي تخضع لتوزيعات

* استاذ/قسم الاحصاء/كلية الادارة والاقتصاد/الجامعة المستنصرية

غير طبيعية [2] . إلا أن محاكاة الظواهر ذات المتغيرات المتعددة التي لا تخضع للتوزيع الطبيعي ليست شائعة بسبب عدم توفر الخوارزميات التي تجعل من ذلك ممكناً [3] .

إن أحد ابرز التوزيعات شائعة الاستخدام في الجوانب النظرية والتطبيقية سواء أكان أحادي أو متعدد للمتغيرات ، يتمثل بتوزيع بيتا ، إذ يقال بأن للمتغير y توزيع بيتا بالمعلمتين a و b ، إذا امتلك دالة كثافة الاحتمال :

$$f_Y(y)=\frac{y^{a-1}(1-y)^{b-1}}{\mathrm{B}(a,b)} \quad , \quad 0<y<1 \quad , \quad a,b>0 \quad , \quad ----(1)$$

حيث أن $\mathrm{B}(a,b)=\int_0^1 y^{a-1}(1-y)^{b-1}\,dy$ ، عبارة عن دالة بيتا الرياضية المعروفة .

أما لو افترضنا أن عينة عشوائية بحجم n قد سحبت من مجتمع يخضع للتوزيع الطبيعي ذي m من المتغيرات ، بحيث أن $m\le n$ ، بمتجه المتوسطات $\underline{\mu}$ ومصفوفة تباين وتباين مشترك Σ ، فان مقدر مصفوفة التباين والتباين المشترك :

$$S=\frac{\sum_{i=1}^{n}(\underline{x}_i-\bar{\underline{x}})(\underline{x}_i-\bar{\underline{x}})'}{n-1}=\frac{A}{n-1} \qquad ----(2)$$

سيخضع لتوزيع وشرت (Wishart distribution) ، ب $r=n-1$ من درجات الحرية و $V=\Sigma/r$ كمصفوفة متوسطات ، وعلى وفق دالة كثافة الاحتمال [1] ،

$$f_S(s)=\begin{cases}\dfrac{|s|^{\frac{r-m-1}{2}}\,e^{-\frac{1}{2}tr\,V^{-1}s}}{2^{\frac{rm}{2}}\,\pi^{\frac{m(m-1)}{4}}\,|V|^{\frac{r}{2}}\prod_{i=1}^{m}\Gamma\left(\frac{r+1-i}{2}\right)} & , \text{if } s \text{ is positivedefinite} \\ 0 & , \text{o.w.}\end{cases} \qquad ----(3)$$

إذ يشار في الأدبيات العلمية لما ورد أعلاه بالشكل $s \sim W_m(r, V)$.

فلو كان لدينا $s_j \sim W_m(r_j, V)$ ، $(j=1,2)$ ، حيث أن s_1 و s_2 مستقلتين عن بعضهما البعض ، فأنه سيقال بأن للمصفوفة $H = C^{-1} S_2 \left(C^{-1}\right)'$ ، توزيع بيتا متعدد المتغيرات (Multivariate Beta distribution) ، وبدالة كثافة الاحتمال [1] :

$$f_H(h) = \begin{cases} \dfrac{\Gamma_m\left(\dfrac{r_1+r_2}{2}\right)}{\Gamma_m\left(\dfrac{r_1}{2}\right)\Gamma_m\left(\dfrac{r_2}{2}\right)} |h|^{(r_2-m-1)/2} |I-h|^{(r_1-m-1)/2} & , \quad h>0 \ , I-h>0 \\ 0 & , \quad o.w. \end{cases} \qquad ----(4)$$

حيث أن $CC' = S_1 + S_2$ ، وأن C ، هي مصفوفة مثلث سفلي .

وبناء" على ما ورد أعلاه ، ولعدم توفر الخوارزميات [3] ، التي تجعل من محاكاة الظواهر المستندة على متغيرات متعدده أمرا" ممكنا" ، فقد آثرنا هنا تقديم خوارزمية عامة لتوليد بيانات تخضع لتوزيع بيتا متعدد المتغيرات ، ليمثل هدف بحثنا هذا .

II. الطريقة الخليطة [2,3] The Mixture Method

إذا افترضنا بأن للمتغير العشوائي K دالة كثافة احتمال (إذا كان المتغير K مستمرا") ، أو دالة كتلة احتمال (إذا كان لمتغير العشوائي K متقطعا") ، تتمثل بالشكل :

$$u(k;\theta) \qquad ----(5)$$

على أن θ يمكن أن تكون معلمة واحدة أو متجه من المعالم ، فأنه بالاعتماد على $K = k$ ، يمكن توليد m من المتغيرات $x_1, x_2, ..., x_m$ ، المستقلة عن بعضها البعض ، وكل منها يمتلك دالة كثافة احتمال شرطية (أو دالة كتلة احتمال شرطية) ،

هي عبارة عن :

$$p(x_i|k,\eta) \quad , i=1,2,...,m \qquad ----(6)$$

حيث أن η يمكن أن تكون هي الأخرى معلمة واحدة أو متجه من المعالم .

وبضرب المعادلتين في (5) و (6) مع بعضهما البعض ، سنحصل على دالة كثافة (كتلة) احتمال مشتركة ل $x_1,x_2,...,x_m$ و K ، بالشكل :

$$g(x_1,x_2,...,x_m,k,\theta,\eta)=u(k,\theta)\prod_{i=1}^{m}p(x_i|k,\eta) \qquad ----(7)$$

لقد لاحظ Minhajuddin و Harris و Schucany عام 2004 [3] ، بأن دالة كثافة (أو كتلة) الاحتمال المشتركة الواردة في المعادلة (7) متماثلة (symmetric) بالنسبة لأي متغيرين مثل x_i و x_j ، حيث أن $i \neq j$ ، وأن $i,j=1,2,...,m$ ، فاستنتجوا بأن المتغيرات العشوائية $x_1,x_2,...,x_m$ ، قابلة للتبادل (Exchangable) ، وأنه بالاعتماد على $K=k$ ، فأن x_i و x_j سيكونا مستقلين ، وأنه يمكن الحصول على دالة كثافة (أو كتلة) الاحتمال المشتركة للمتغيرات $x_1,x_2,...,x_m$ بالشكل ،

$$f(x_1,x_2,...,x_m;\theta,\eta)=\begin{cases}\int_{\forall k} g(x_1,x_2,...,x_m,k;\theta,\eta)\,dk & , \text{if } K \text{ continuous}\\ \sum_{\forall k} g(x_1,x_2,...,x_m,k;\theta,\eta) & , \text{if } K \text{ discrete}\end{cases} \qquad ----(8)$$

ان قيمة المعلمة (أو قيمة متجه المعالم) η في الدالة $f(x_1,x_2,...,x_m;\theta,\eta)$ ، تلعب دورا" مهما" جدا" في السيطرة على الارتباط مابين المتغيرات القابلة للتبادل x_i و x_j ($i \neq j$) ، على وفق أقيام المشاهدات التي تخضع للمتغير K .

وبناء" على الفلسفة الواردة في أعلاه ، خلص كلٌ من Minhajuddin وHarris و Schucany عام 2004 [3] ، الى الخطوتين التالية لمحاكاة التوزيعات ذوات المتغيرات المتعددة.

1) محاكاة المتغير $K = k$ على وفق دالة كثافة (كتلة) الاحتمال الحدية $u(k;\theta)$.

2) اعتمادا" على كل قيمة مولدة للمتغير $K = k$ ، يتم توليد m من الاقيام المستقلة للمتغيرات x_i ($i=1,2,...,m$) ، قيمة لكل متغير ، وذلك من خلال التوزيع اللاحق $p(x_i \mid K = k, \eta)$.

إن احد ابرز إفرازات نظرية بيز ، قد تمثل بعوائل التوزيعات الأولية المترافقة ، فعائلة توزيع بيتا مثلا" هي المرافقة لعائلة توزيع ذي الحدين ، بمعنى أنه لو كان X يخضع لتوزيع بيتا بالمعلمتين α و β ، وأن $K \mid X = x$ يخضع لتوزيع ذي الحدين ، فان هذا سيؤدي إلى خضوع K لتوزيع حدي (غير معتمد) هو عبارة عن توزيع بيتا – ذي الحدين (Beta-Binomial) ، كما أن عائلة توزيع گاما هي المرافقة لعائلة توزيع بواسون ، وهذا يؤدي إلى خضوع K لتوزيع حدي (غير معتمد) هو عبارة عن توزيع ذي الحدين السالب .

لقد استفاد منظروا الطريقة الخليطة قيد الدراسة من إفراز نظرية بيز المذكور أعلاه ، من خلال البدء بالتوزيع الحدي (غير المعتمد) للمتغير K ، وبالاعتماد على القيمة المولدة $K = k$ ، تم توليد أقيام المتغيرات $x_1, x_2, ..., x_m$ ، ولكل على وفق العائلة المترافقة معها .

III. محاكاة عائلة بيتا متعدد المتغيرات

يذكر كلٌ من Patil و Boswell و Joshi و Ratnaparkhi عام 1984 [5]، الصيغة التالية لتوزيع بيتا – ذي الحدين :

$$p(k) = \frac{C_k^{-\alpha}\, C_{v-k}^{-\beta}}{C_v^{-\alpha-\beta}} \quad , \quad k = 0,1,...,v \quad , \quad v = 1,2,... \qquad ---(9)$$

وباستخدام العلاقة الرياضية [4] ، $C_{d_1}^{-d_2} = (-1)^{d_1} C_{d_1}^{d_1+d_2-1}$ ، فانه يمكن كتابة $p(k)$ في المعادلة (9) بالشكل :

$$p(k) = \frac{(-1)^k C_k^{\alpha+k-1} (-1)^{v-k} C_{v-k}^{v-k+\beta-1}}{(-1)^v C_v^{v+\alpha+\beta-1}}$$

وبالاختصار وإعادة كتابة التوافيق بشكل مفكوكات ، يكون ،

$$\begin{aligned} p(k) &= \frac{v!(\alpha+k-1)!(v-k+\beta-1)!(\alpha+\beta-1)!}{k!(\alpha-1)!(\beta-1)!(v-k)!(v+\alpha+\beta-1)!} \\ &= C_k^v \frac{\Gamma(k+\alpha)\Gamma(v+\beta-k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(v+\alpha+\beta)} \\ &= C_k^v \frac{B(k+v, v+\beta-k)}{B(\alpha,\beta)} \qquad ----(10) \end{aligned}$$

والصيغة المذكورة في (10) أعلاه هي شكل آخر لدالة كتلة احتمال توزيع بيتا – ذي الحدين .

إن متوسط وتباين المتغير K سيكونا عبارة عن [5] :

$$E(K) = \frac{v\alpha}{\alpha+\beta} \quad , \quad Var(K) = \frac{v\alpha\beta(\alpha+\beta+v)}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad ----(11)$$

ومن ذلك فأن ،

$$E(K^2) = Var(K) - [E(K)]^2 = \frac{v\alpha(v\alpha+v+\beta)}{(\alpha+\beta+1)(\alpha+\beta)} \qquad ----(12)$$

وتأسيسا" على ما ورد أعلاه وما ورد في المعادلة (7) ، فان دالة الكثافة الاحتمالية المشتركة مابين أي من المتغيرات x_i $(i=1,2,...,m)$ والمتغير K ستكون :

$$\begin{aligned} g(x_i, k; v, \alpha, \beta) &= \frac{x_i^{\alpha+k-1}(1-x_i)^{v+\beta-k-1}}{B(k+\alpha, v+\beta-k)} \cdot \frac{C_k^v B(k+\alpha, v+\beta-k)}{B(\alpha,\beta)} \\ &= \frac{x_i^{\alpha-1}(1-x_i)^{\beta-1}}{B(\alpha,\beta)} . C_k^v x_i^k (1-x_i)^{v-k} \qquad ----(13) \end{aligned}$$

إذ أن $x_i | k$ $(i=1,2,...,m)$ ، يخضع لتوزيع بيتا بالمعلمتين $\alpha+k$ و $v+\beta-k$.

وبأخذ المجموع لكل قيم K ، حيث أن $k=0,1,2...,v$ ، يمكن الحصول على دالة الاحتمال الحدية للمتغير x_i ، إذ أن الحدود المتضمنة المتغير K ، في المعادلة (13) ، عبارة عن دالة ذي الحدين بالمعلمتين v و x_i ، ومجموعها يؤول للواحد الصحيح ، فيكون توزيع أي متغير مثل $x_i\,(i=1,2,...,m)$ ، عبارة عن توزيع بيتا بالمعلمتين α و β ، وعليه فأن متوسط وتباين أي متغير مثل x_i سيكونا عبارة عن :

$$E(x_i)=\frac{\alpha}{\alpha+\beta} \quad , \quad Var(x_i)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad ----(14)$$

وبما أنه يتم توليد قيمة لكل متغير من المتغيرات $x_1,x_2,...,x_m$ بشكل مستقل عن القيمة المولدة لمتغير آخر ، اعتمادا" على نفس القيمة للمتغير K ، كما أسلفنا ، فأنه استنادا" لمعادلتين (13) و (7) ، يكون التوزيع الشرطي لكل متغير مثل $x_i\ (i=1,2,...,m)$ اعتمادا" على $K=k$ ، هو عبارة عن توزيع بيتا بالمعلمتين $k+\alpha$ و $v+\beta-k$ ، وبدالة كثافة الاحتمال :

$$p\left(x_i \mid k\right)=\frac{x_i^{k+\alpha-1}(1-x_i)^{v+\beta-k-1}}{\mathrm{B}(k+\alpha,v+\beta-k)} \quad , \ i=1,2,...,m \quad , \quad ----(15)$$

وعلى ذلك فان :

$$\left.\begin{aligned} E_{X_i|K}(X_i)&=\frac{K+\alpha}{\alpha+\beta+v} \\ Var_{X_i|K}(X_i)&=\frac{(K+\alpha)(\beta+v-K)}{(\alpha+\beta+v+1)(\alpha+\beta+v)^2} \end{aligned}\right\} \qquad ----(16)$$

وأنه لكل $1\le i\ne j\le m$ ، يكون :

$$\begin{aligned} E(X_iX_j)&=E_K\left[E_{X_iX_j|K}(X_iX_j)\right] \\ &=E_K\left[E_{X_i|K}(X_i).E_{X_j|K}(X_j)\right] \\ &=E_K\left[\frac{(K+\alpha)^2}{(\alpha+\beta+v)^2}\right] \\ &=E_K\left[\frac{K^2+2\alpha K+\alpha^2}{(\alpha+\beta+v)^2}\right] \end{aligned}$$

وبالتعويض من المعادلتين (11) و (12) ، وجملة من العمليات البسيطة ، يكون :

$$E(X_i X_j) = \frac{\alpha\{v(\alpha+1)+\alpha(\alpha+\beta+1)\}}{(\alpha+\beta+v)(\alpha+\beta+1)(\alpha+\beta)}$$

فتكون قيمة الارتباط مابين أي متغيرين مثل x_i و x_j ($i \neq j$) عبارة عن :

$$\rho_{X_i X_j} = \frac{E(X_i X_j) - E(X_i)E(X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

$$= \frac{v\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)(\alpha+\beta+v)} \cdot \frac{(\alpha+\beta)^2(\alpha+\beta+1)}{\alpha\beta}$$

$$= \frac{v}{(\alpha+\beta+v)} \qquad ----(17)$$

إن قيمة الارتباط في (17) أعلاه ، ذات أهمية بالغة في عمل الخوارزمية العامة قيد الطرح ، إذ من خلالها يمكن تحديد قيمة المعلمة v في توزيع بيتا – ذي الحدين ، حيث يكون :

$$v = \frac{(\alpha+\beta)\rho}{1-\rho} \qquad ----(18)$$

فتدخل قيمة الارتباط ما بين المتغيرات في عمل الخوارزمية على وفق هذا النحو ، لما للارتباط من أهمية بالغة في تحليل العلاقات مابين المتغيرات المتعددة .

IV- خوارزمية المحاكاة وتنفيذها

بعد ما ورد في فقرات البحث السابقة ، وعلى الأخص الفقرة (III) ، يمكن تلخيص خوارزمية المحاكاة ، إجمالاً بالآتي :

a) يتم توليد قيمة p ، لتخضع لتوزيع بيتا بالمعلمتين α و β ، وعلى وفق دالة كثافة الاحتمال في (1) ، إذ أن :

$\alpha = a$ و $\beta = b$.

b) توليد قيمة k من خلال خوارزمية توزيع ذي الحدين ، على اعتبار قيمة p المولدة في الخطوة (a) أعلاه كقيمة معلمة لهذا التوزيع جنبا" إلى جنب مع قيمة v المستخرجة من المعادلة (18) ، كقيمة معلمة هي الأخرى ، لتخضع k بالنتيجة لتوزيع بيتا – ذي الحدين بالمعلمات v و $\alpha^* = \alpha + \beta$ و $\beta^* = \beta$.

c) توليد m من القيم $x_1, x_2, ..., x_m$ ، على أساس قيمة k المولدة في الخطوة (b) السابقة ، وذلك على وفق توزيع بيتا بالمعلمتين $\alpha^* + k$ و $v + \beta^* - k$ ، لتندرج كل قيمة تحت المتغير المناظر لترميزها $X_1, X_2, ..., X_m$ على التوالي كقيمة مولدة على وفقه .

d) يتم تكرار الخطوات السابقة بقدر عدد المشاهدات المراد توليده .

أما البرنامج المكتوب من قبل الباحث بلغة فجوال بيزك الذي يعبر عن الخوارزمية أعلاه ويجعلها قابلة للتنفيذ فهو الآتي :

```
1 REM " THIS PROGRAM WRITTEN BY (DR. SALAH HAMZA
ABID)    "
2 REM "    TO GENERATE OBSERVATIONS FELLOW
MULTIVARIATE      "
3 REM "    BETA DISTRIBUTION WITH CORRELATION
COEFFICENT      "
4 REM "    (rw) AND (m) OF VARIABLES AND (n) OF
OBSERVATIONS    "
5 REM "
**********************************************************"
10 INPUT "rw=";RW
20 INPUT "m=";M
30 INPUT "n=";N
40 INPUT "a=";A
50 INPUT "b=";B
```

```
60 DIM X(M,N)
70 FOR J=1 TO N:RANDOMIZE TIMER:SSA=0:SSB=0:k=0
80 FOR I=1 TO INT(A)
90 SSA=SSA-LOG(RND):NEXT
100 SSA=SSA-(LOG(RND))*(A-INT(A))
110 FOR I=1 TO INT(B):SSB=SSB-LOG(RND):NEXT
120 SSB=SSB-(LOG(RND))*(B-INT(B))
130 P=SSA/(SSA+SSB)
140 V=((A+B)*RW)/(1-RW)
150 FOR I=1 TO INT(V)
160 IF RND <= P THEN K=K+1
170 NEXT
180 IF ((V-INT(V))*RND) <= P THEN K=K+1
190 FOR I=1 TO M:SS1=0:SS2=0
200 FOR L1=1 TO INT(K+A+B):SS1=SS1-LOG(RND):NEXT
210 SS1=SS1-(LOG(RND))*((K+A+B)-INT(K+A+B))
220 FOR L1=1 TO INT(V+A-k):SS2=SS2-LOG(RND):NEXT
230 SS2=SS2-(LOG(RND))*((V+A-k)-INT(V+A-k))
240 X(I,J)=SS1/(SS2+SS1):NEXT:NEXT
250 FOR I=1 TO M:FOR J=1 TO N:PRINT USING "
#.####";X(I,J);:NEXT:PRINT :NEXT
```

إن نتيجة تنفيذ البرنامج أعلاه عند قيم $\rho = 2/3$ و $\alpha = 1$ و $\beta = 2$ و $m = 4$ و $n = 10$ مثلاً ، هي المشاهدات المولدة التالية ، التي تخضع لتوزيع بيتا ذي الاربع متغيرات :

0.4601 0.7486 0.8640 0.2959 0.4480 0.5358 0.6322 0.8494 0.2820
0.7333 0.3218 0.8000 0.5599 0.6305 0.3129 0.6097 0.8640 0.7744
0.3868 0.5131 0.4664 0.9860 0.7510 0.3344 0.8437 0.5272 0.8524
0.9045 0.6982 0.4870 0.6001 0.8169 0.2160 0.4756 0.7563 0.5507
0.7026 0.7920 0.3807 0.4606

المصادر

1) Johnson, N. and Kotz, S. (1978) "Distributions in statistics : continuous multivariate Distributions" , wiley series , USA.

2) Michael, J. and Schucany, W. (2002) "The mixture approach for simulating bivariate Distributions with specific correlations" , The American Statisticians , 56 , pp.48-54 .

3) Minhajuddin, A. & Harris, I. and Schucany, W. (2004) "Simulating multivariate Distributions with specific correlations ", The American Statisticians , 58 , pp.86-98 .

4) Mood, A. & Graybill, F. and Boes, D. (1989) "Introduction to the theory of statistics" , third edition , McGraw Hill company , USA .

5) Patil , G. & Boswell , M. & Joshi , S. and Ratnaparkhi , M. (1984) "Dictionary of statistical distributions " , McGraw Hill company , USA .

.....

.....

.....